LLNL HPC Update

Presented to ASC PI Meeting

Matt Leininger

Lawrence Livermore National Laboratory

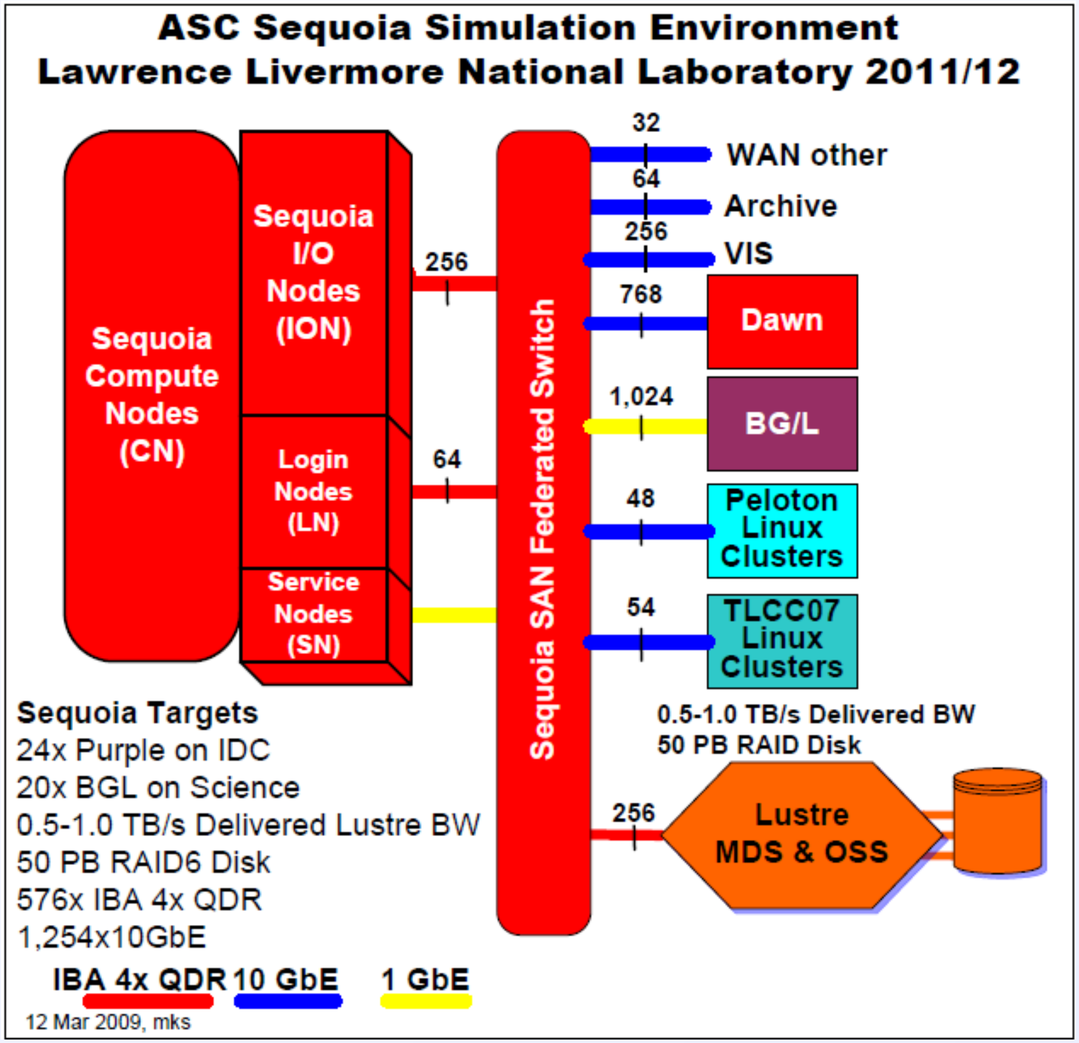
10 February 2010

LLNL-PRES-424187

Talk Overview

- Progress on Sequoia
- Hyperion Data Intensive Testbed
- Lustre operational improvements
- Next generation tri-Laboratory capacity procurement now under way

Sequoia Hierarchal Hardware Architecture in Integrated Simulation Environment



ASC Sequoia Simulation Environment
Lawrence Livermore National Laboratory 2011/12

Sequoia Targets
24x Purple on IDC
20x BGL on Science
0.5-1.0 TB/s Delivered Lustre BW
50 PB RAID6 Disk
576x IBA 4x QDR
1,254x10GbE

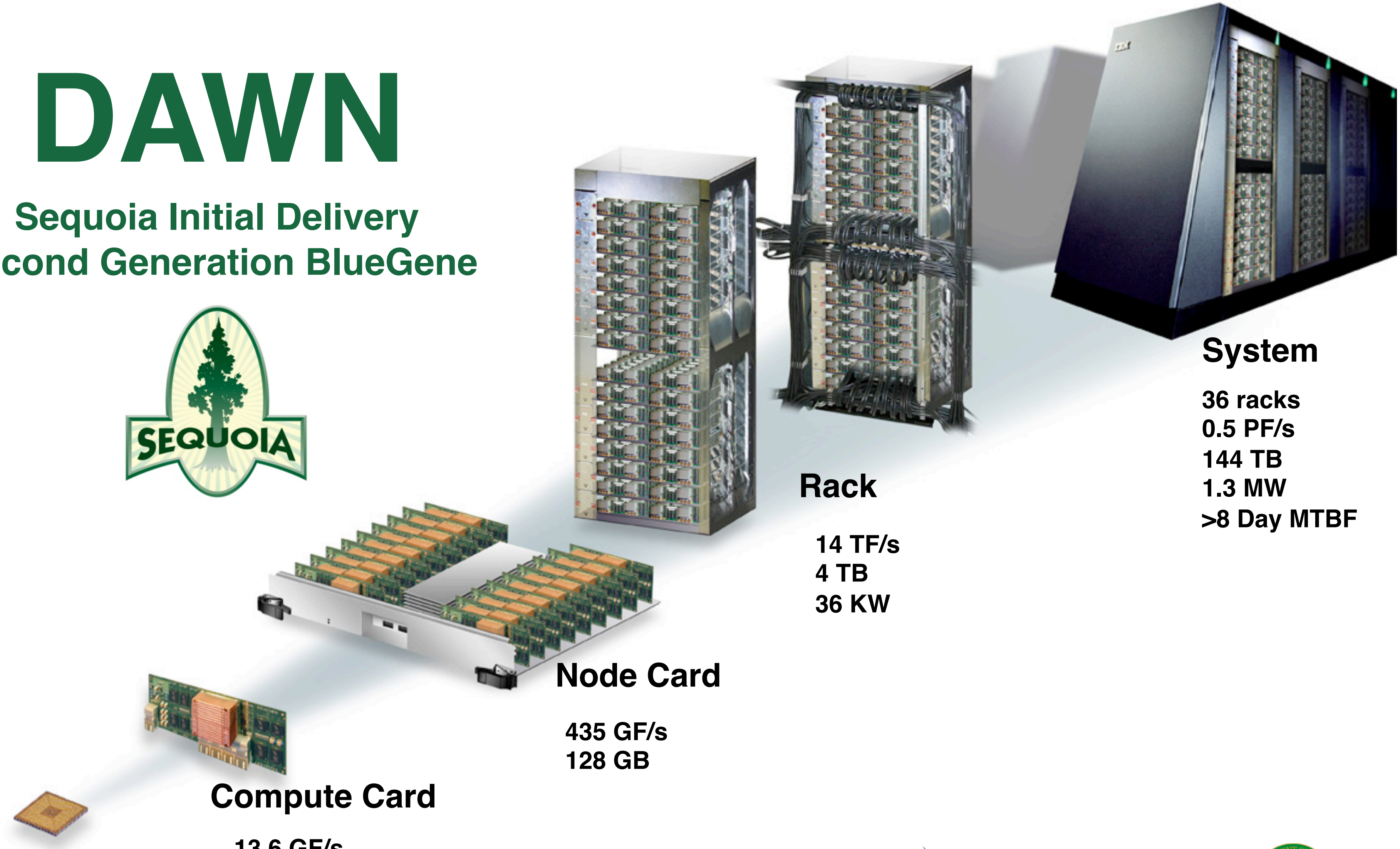
IBA 4x QDR 🟥 10 GbE 🟦 1 GbE 🟨

12 Mar 2009, mks

- Sequoia Statistics
  - 20 PF/s target
  - Memory 1.6 PB, 4 PB/s BW
  - 1.5M Cores
  - 3 PB/s Link BW
  - 60 TB/s bi-section BW
  - 0.5-1.0 TB/s Lustre BW
  - 50 PB Disk
- 8.0MW Power, 3,500 ft^2
- Third generation IBM BlueGene
- Challenges
  - Hardware Scalability
  - Software Scalability
  - Applications Scalability

DAWN

Sequoia Initial Delivery
Second Generation BlueGene



Chip

850 MHz PPC 450
4 cores/4 threads
13.6 GF/s Peak
8 MB EDRAM

Compute Card

13.6 GF/s
4.0 GB DDR2
13.6 GB/s Memory BW
0.75 GB/s 3D Torus BW

Node Card

435 GF/s
128 GB

Rack

14 TF/s
4 TB
36 KW

System

36 racks
0.5 PF/s
144 TB
1.3 MW
>8 Day MTBF

Dawn now in Classified Service and delivering to the program

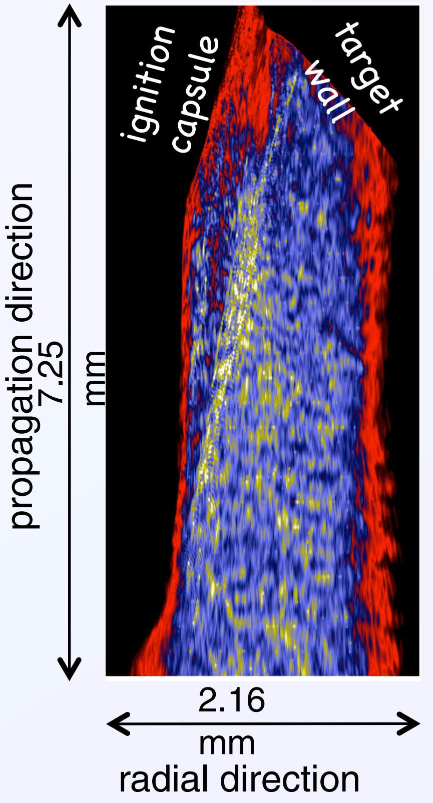
- Dawn hardware delivery started 19 Jan 2009. Rapid deployment of 36 racks completed ahead of an aggressive schedule
- Full Synthetic Workload acceptance test successfully completed 26 March 2009
- Twelve codes from Tri-Lab community ran on system during science runs
- Dawn Dedication 27 May 2009
- Now in classified service

The first half of DAWN (initial delivery of Sequoia) was received at the TerascaleSimulation Facility in late January, 2009

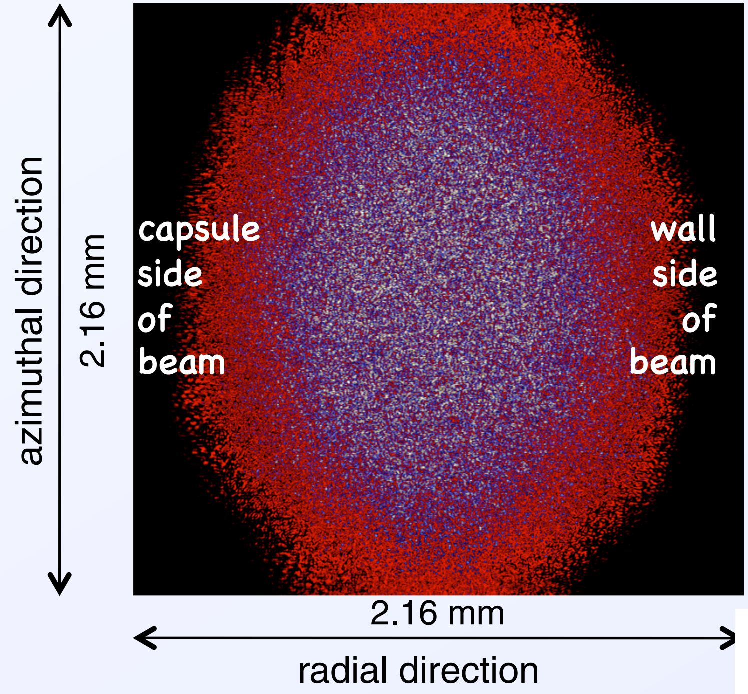
As an example of the interdependence of theory and experiment, NIF recently simulated an entire 30^0 beam quad with improved physics in preparation for Ignition

Ignition Design
30^0 Quad Intensity (W/cm^2)

Quad Propagation

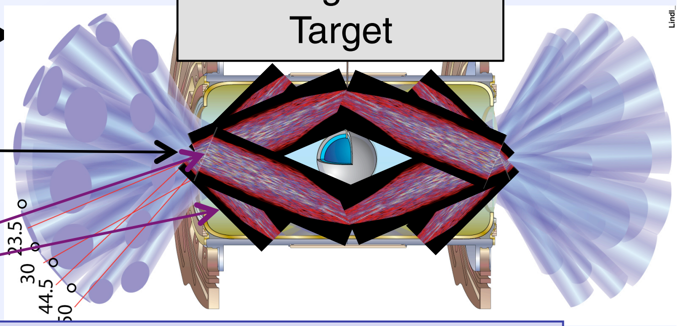


Quad Cross Section



- our simulations:
  - -- resolve laser speckles
  - -- include improved physics
  - -- “more of the problem”
  - -- show 9.5% reflectivity (within spec)

NIF Ignition Target



- The quad of beams we simulated is in this cone
- We plan to simulate two crossing quads later this year

A mammoth four-week calculation completed June 10 using all of the 500 TF Dawn to support first ignition experiments…

2009 National Medals of Science and of Technology recognize LLNL accomplishments and collaborations

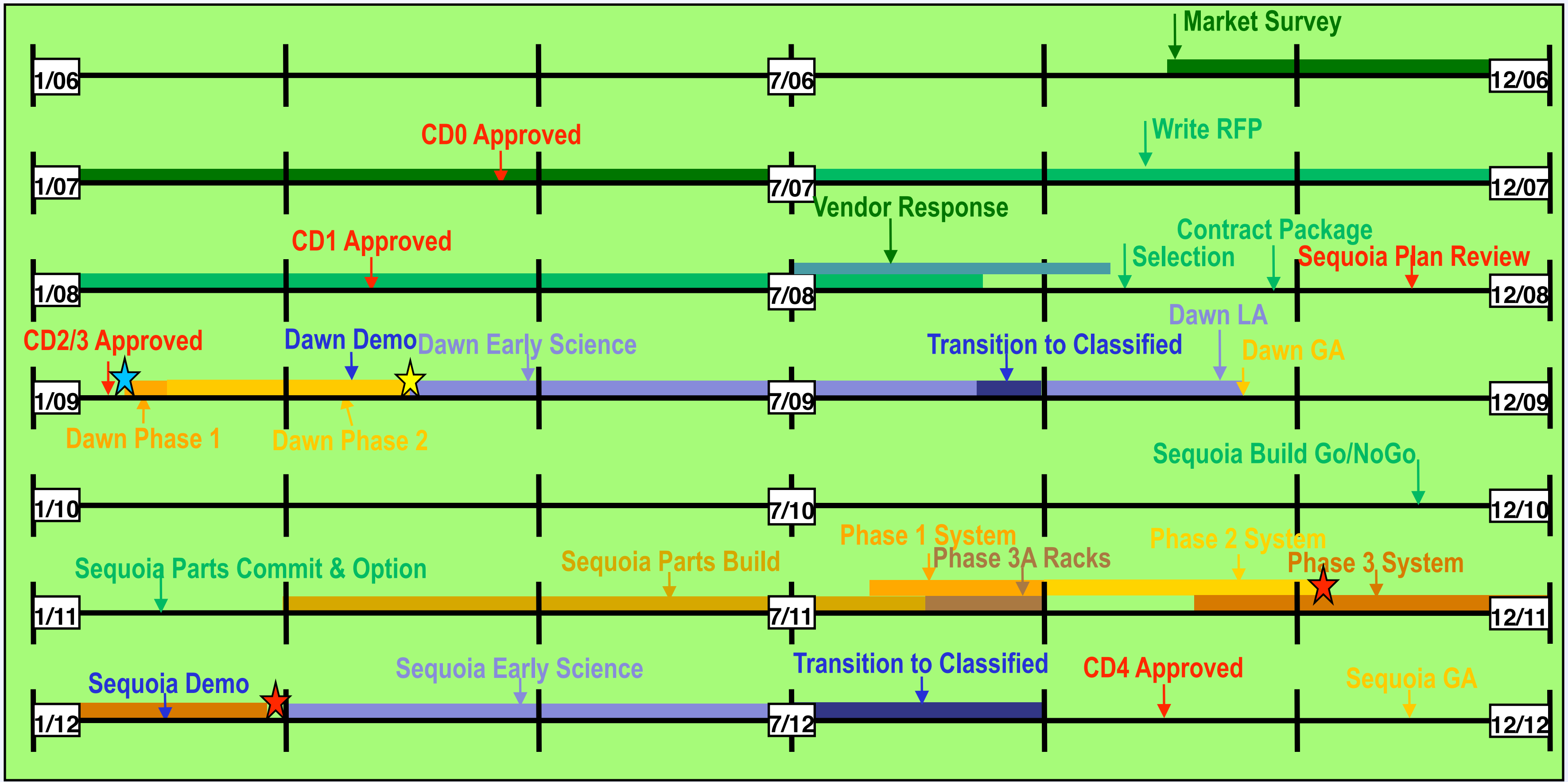
- **Berni Alder, computational pioneer**
  - Founder of molecular dynamics
  - Recognized for large-scale simulations to solve quantum mechanics problems
- **IBM - Blue Gene**
  - Series of energy-efficient supercomputers
  - LLNL and ANL partnership strongly impacted extreme-scale design and DOE supported IBM R&D

President Obama presented the Medals to Berni Alder and Sam Palmisano (IBM CEO) at a White House Ceremony on October 7, 2009

Recent Progress Developing Sequoia

- RIT1 release to fabrication
  - Verification of innovative SMP hardware enhancements continuing
  - Prototype fabrication delayed 2.5 months
- Updated power estimates
- Compiler and tools progress with SMP programming models
- Hardware and software simulators being used to evaluate performance
- System software work continuing
  - Innovative SIOD model architected and implemented
  - RHEL for ION, LN and SN almost complete
  - OS Booting and stress test running on hw simulator
- Projecting GO/NOGO decision in October 2010

Sequoia Timeline Delivers Petascale Resources to the Program



Sequoia Five Years Planned Lifetime Through CY17

☆ Sequoia contract award ☆ Dawn system acceptance ★ Sequoia phase 2 & final system acceptance

Sequoia reduction strategy is multifaceted and intended to provide a production multi-petaflop simulation environment

Purple and BG/L pedigree leveraged by NNSA User Facility

Dawn

IBM BG/P

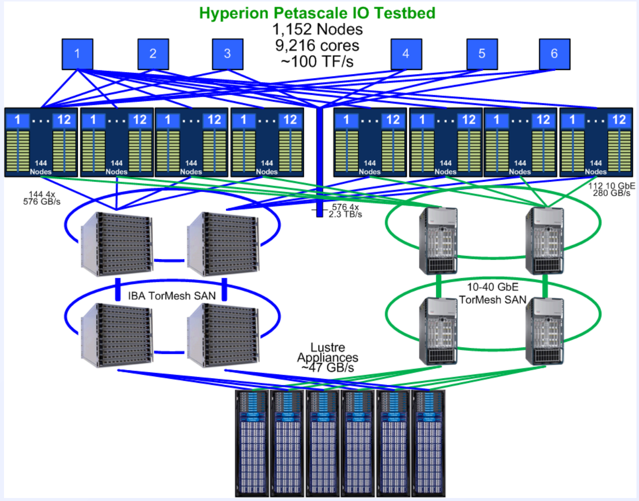
Weapons codes transition to multicore architecture 2009-2012

SEQUOIA

IBM BG/Q

Innovative cost-shared test bed, "Hyperion" to prepare for the massive I/O demand (1+TB/s)

Cost-sharing model supported by LLNL Industrial Partnerships Organization (IPO)



Hyperion Partnership Update

- 2009 HPCWire Award for best “Government & Industry Partnership”
- Moved system to green network and have foreign national collaborators on the machine
- IBM/Houston considering joining partnership
  - Test next release of HPSS at scale
  - Collaborators (Sun) to test Lustre HSM back-end
- Collaboration wants to develop outreach activity to ISV community
- Major IO expansion planned for FY10 for scale testing in preparation for Sequoia

Lustre File Systems Grow To Meet Computing Demands

- OCF file systems expanded by 29%
  - Added /p/lscratchd, a 20GB/s, 755 TB file system to support Coastal.
- SCF file systems expanded by 143%
  - Added /p/lscratch2, a 70 GB/s, 2.7 PB file system to support Dawn.
  - Added /p/lscratch4, a 60 GB/s, 2.3 PB file system to meet SCF capacity computational demands.
- Operations improvements using unified Storage Scalable Unit (SSU)
  - Upgraded the Metadata Servers center-wide
    - Now have identical MDS configuration for all file systems
    - Improved performance by XX%
  - Reworked oldest file system to modern Storage Scalable Unit
    - Lustre servers without local disk for OS
    - Use n+2 parity for Lustre bulk Object Store

Lustre operational improvements over the last year have enable efficiency and improved reliability

- Software and Development:
  - Implemented Failover on our servers
  - Worked extensively with Sun to align our releases
- Operations
  - Extensive training for Operators
  - Leveraged System Admin Group to spread Lustre knowledge.
- Software and Operational efforts reduced off-hours support calls by 50%!

Tri-Laboratory Capacity Cluster (TLCC11) Overview

- Second step in Tri-lab capacity cluster for ASC Program
- Technical
  - Deliver working clusters of multiple sizes to the Program
  - Purchase cost effective balanced commodity SU’s
  - Define SU config with room for vendor innovation and differentiation
  - Receiving site works with vendor to aggregate SU’s
- Financial
  - Investment for GFY11+12
  - ~$29 - $50M to purchase SU’s ($29M + options)
  - Single contract with multiple delivery sites
  - SU allocations determined by HQ
  - Provide vendor with flexibility to optimize build+ delivery
  - Long term partnership via shared risk

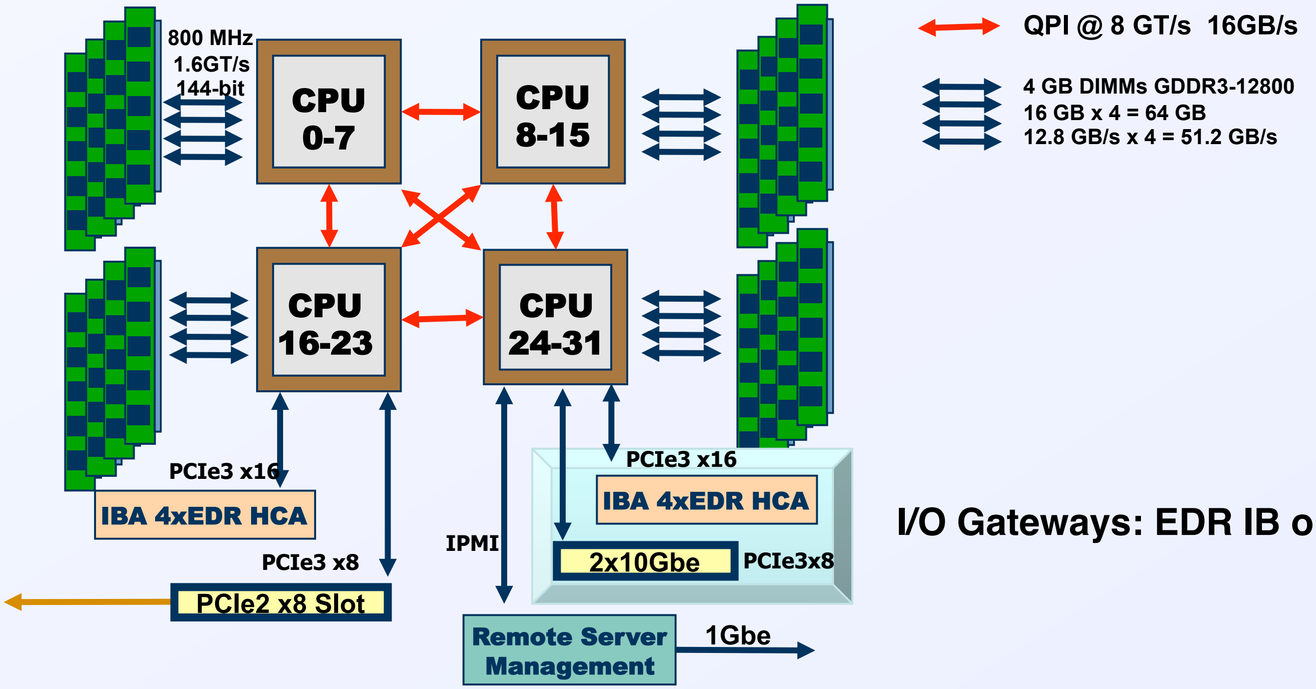
TLCC11 Procurement

- Procurement Strategy:
  - Define consistent SU for duration of contract with room for configuration options (memory, clock speed, network, rack power)
  - Commit to volume purchase over two years, let vendor optimize supply chain
  - Shared risk model for forward pricing of commodities
  - Reduce site and applications support costs through Common Computing (hardware and software) Environment

- Follow TLCC07 build model:
  - Standard SU (HW+SW) with options
  - Vendor Partner builds SU with pre-ship test
  - Vendor Partner delivers SU to site with post-ship test
  - Local site and Vendor Partner aggregate SU's on site

Major focus on delivering capacity to the program while reducing overall Total Cost of Ownership (TCO)

Example TLCC-11 Node Architecture

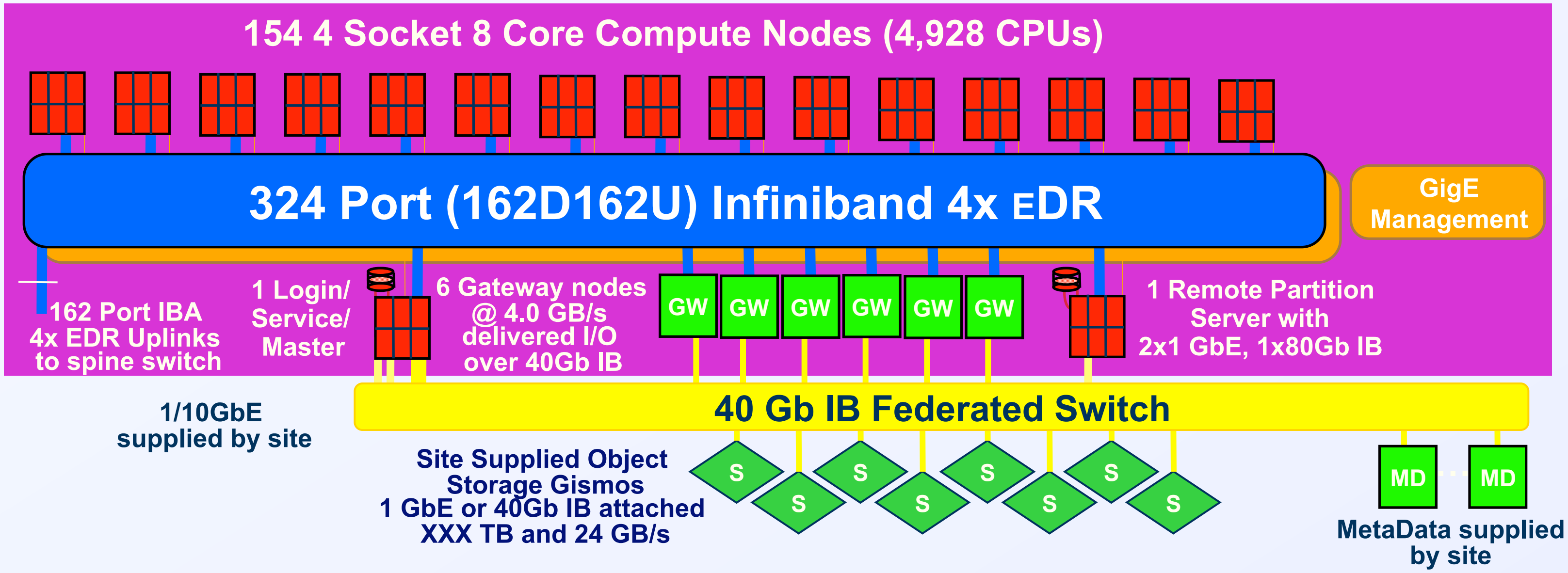


I/O Gateways: EDR IB or 10Gbe

2 or 4 or more socket nodes available

- •6-12 core at 2.0-3.0 GHz (95W)
- •4-8 GB DDR3-12800 DIMM, 1-2 slots per channel
- •Node peak is 384 – 576 GF/s
- •64-256 GB memory (2 – 5 GB/core)
- •51.2 GB/s memory BW per socket and 204.8 per node
- •12.5+12.5 GB/s IBA 4x EDR BW

Target SU Configuration



154 4 Socket 8 Core Compute Nodes (4,928 CPUs)

324 Port (162D162U) Infiniband 4x EDR

GigE Management

162 Port IBA 4x EDR Uplinks to spine switch

1 Login/ Service/ Master

6 Gateway nodes @ 4.0 GB/s delivered I/O over 40Gb IB

GW GW GW GW GW GW

1 Remote Partition Server with 2x1 GbE, 1x80Gb IB

1/10GbE supplied by site

40 Gb IB Federated Switch

Site Supplied Object Storage Gismos 1 GbE or 40Gb IB attached XXX TB and 24 GB/s

S S S S S S S S

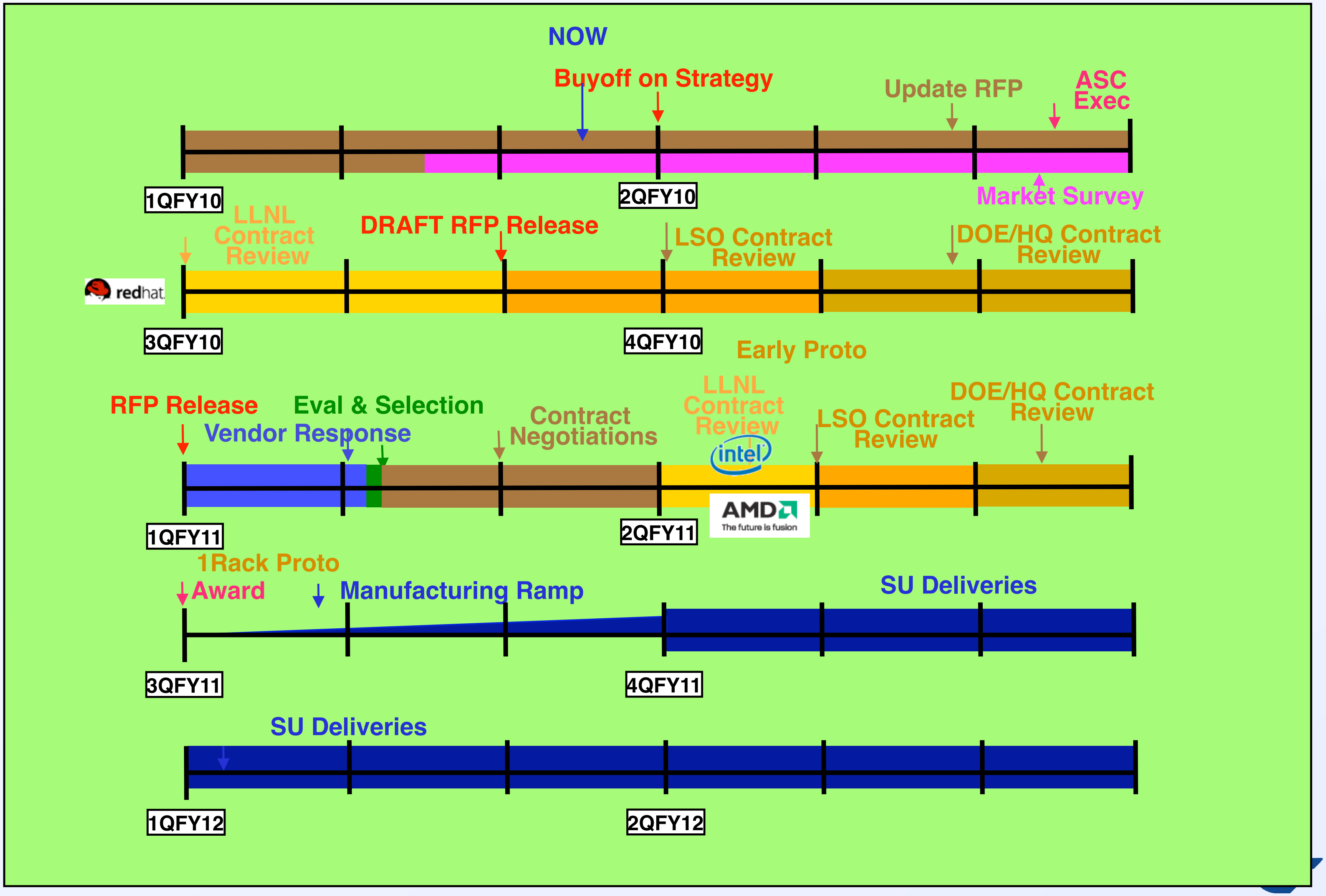
MD MD

MetaData supplied by site

System Parameters: 162 nodes and 62-93 TF/s per SU

- 384-576 GF/s quad socket 3.0 GHz 6-12 core x86_64 (95W) SMP nodes
- 64-256 GB DDR3 SDRAM
- <3 μs, 25 GB/s MPI latency and Bandwidth over IBA 4x EDR
  - Built from 36-port switches for 4 and 8 SU aggregate cluster configurations
- Support 10 GB/s transfers to Archive over IBA links from Login node.
- No local disk. Remote boot and SRP target for root and swap partitions on RAID5 device for improved RAS
- 4 GB/s POSIX serial I/O to any file system
- IO Bandwidth 24 GB/s delivered parallel I/O performance
- Software for build and acceptance RHEL6, Moab/SLURM, OpenFabrics, MPICH2/OpenMPI, GNU Fortran, C and C++ compiler , Commercial Fortran compiler.

TLCC11 Timeline



NOW

Buyoff on Strategy

Update RFP

ASC Exec

1QFY10

2QFY10

Market Survey

LLNL Contract Review

DRAFT RFP Release

LSO Contract Review

DOE/HQ Contract Review

3QFY10

4QFY10

Early Proto

RFP Release

Eval & Selection

Vendor Response

Contract Negotiations

LLNL Contract Review

LSO Contract Review

DOE/HQ Contract Review

1QFY11

2QFY11

1Rack Proto

Award

Manufacturing Ramp

SU Deliveries

3QFY11

4QFY11

SU Deliveries

1QFY12

2QFY12

TLCC11 Current Status

- Technical committee with Tri-lab representation
  - Bi-weekly calls
  - Focus is on Market Survey and Draft SOW
- Market Survey:
  - TriLab NDA done - Sent to vendors 2/5
  - Planning a 2 day “brief-fest” 2/22-23 in Livermore
  - Expecting 14 vendors @1 hr each to present
- Draft SOW:
  - Site facilities updates due next week
  - Software: CCE (Common Cluster Environment) text added
  - HW Technical requirements - tbd after market survey

Summary

- Sequoia project has made significant progress in the last year
  - Dawn delivering to the program
  - Sequoia development progressing toward prototype this summer and GO/NOGO in October 2010
- Hyperion project delivering results and will expand to include a data intensive testbed
- TLCC11 procurement underway and progressing
- Lustre operational improvements have significantly streamlined operations and improved reliability
- Multiple LLNL HPC activities won national awards this last year